Person Search by Text Attribute Query as Zero-Shot Learning

Qi Dong
Queen Mary University of London
q.dong@qmul.ac.uk

Shaogang Gong
Queen Mary University of London
s.gong@qmul.ac.uk

Xiatian Zhu
Vision Semantics Ltd.
eddy.zhuxt@gmail.com

Abstract

Existing person search methods predominantly assume the availability of at least one-shot imagery *sample of the queried person. This assumption is limited in circumstances where only a brief* textual *(or verbal) description of the target person is available. In this work, we present a deep learning method for text attribute description based person search without any query imagery. Whilst conventional cross-modality matching methods, such as global visual-textual embedding based zero-shot learning and local individual attribute recognition, are functionally applicable, they are limited by several assumptions invalid to person search in deployment scale, data quality, and/or category name semantics. We overcome these issues by formulating an Attribute-Image Hierarchical Matching (AIHM) model. It is able to more reliably match text attribute descriptions with noisy surveillance person images by jointly learning global category-level and local attribute-level textual-visual embedding as well as matching. Extensive evaluations demonstrate the superiority of our AIHM model over a wide variety of state-of-the-art methods on three publicly available attribute labelled surveillance person search benchmarks: Market-1501, DukeMTMC, and PA100K.*

1. Introduction

Person search in large scale videos is a challenging problem with extensive applications in forensic video analysis and live video surveillance [11]. From increasing numbers of smart cities across the world equipped with tens to hundreds of thousands of 24/7 surveillance cameras per city, a massive quantity of raw video data is cumulatively produced daily. It is infeasible for human operators to *manually* search people (e.g. criminal suspects or missing persons) in such data. Automated person search becomes essential.

Most existing person search methods are based on *image* queries (probes), also known as *person re-identification* [11, 13, 21, 39, 40]. Given a query image, a system computes pairwise visual similarity scores between the query image and every gallery image in the test data. The top ranks with the highest similarity scores are considered as possible matches. Such an operation assumes that at least *one image* (one-shot) of the queried person is available for initiating the search. This is limited when there is only *verbal or text* description of the target persons.

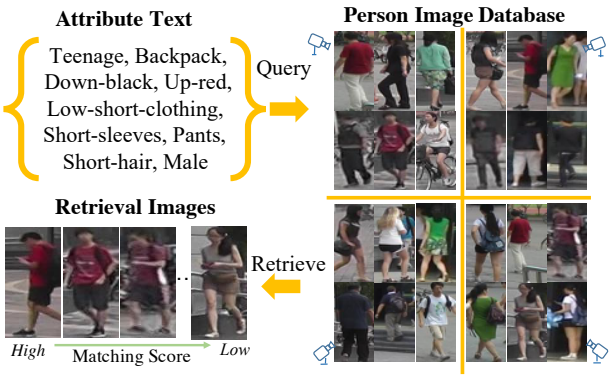

Figure 1: Person search by text attributes (keywords).

There are a number of attempts on person search by text queries, e.g. natural language descriptions [20, 19] or discrete text attributes [37, 16, 32]. To learn such search systems, labelling a large training dataset across textual and visual data modalities is necessary. Elaborative language descriptions not only require more expensive training data labelling, but also present significant computational challenges. This is due to ambiguities in interpretation between language descriptions and image appearance such that: (1) significant and/or subtle visual variations for the same language description; (2) flexible sentence syntax in language descriptions for the same image; and (3) modelling the sequential word dependence in a sentence is a difficult problem, particularly for long descriptions.

In contrast, *text attribute descriptions* are not only much cheaper in collecting labelled training data, but also more tractable in model optimisation. Importantly, they eliminate the need for modelling complex sentence structures and their correlations to the same visual appearance, and vice versa. Whilst giving a compromise of weaker appearance descriptive capacity, using text attributes favourably enables a more robust and computationally tractable means

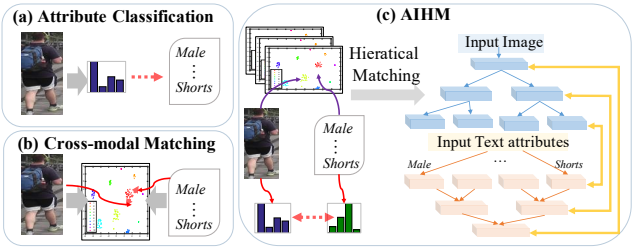

Figure 2: Model architectures for attribute query person search. (a) Individual attribute classification, i.e. *local attribute-level modelling*. (b) Cross-modal matching, i.e. *global category-level modelling*. (c) The proposed attribute-image hierarchical matching (AIHM), integrating both local and global modelling.

to text-query in person search without image probes.

Text attribute query person search is largely understudied in the literature. There exist very few attempts. An intuitive approach is to estimate an attribute vector (text description) of each person image, and then to match the attribute vector of the query person with those of all the gallery person images [16, 32] (Fig 2(a)). By treating the attribute labels *independently*, this method scales flexibly to handling the huge attribute combination space. However, it suffers from lacking a supporting context that accounts for a holistic interpretation of all the text attributes as a whole which helps the text-image matching in person search. The current state-of-the-art model, AAIPR [37] (Fig 2(b)), takes the text-image matching strategy but loses the generalisation scalability of individual attribute modelling.

In this work, for the first time we formulate the problem of text attribute query person search as a zero-shot learning (ZSL) problem [35, 10]. This is because the potential test query categories (text attribute combinations) exist at large scale in reality, but only a small proportion of them can be available for model training due to the high cost for exhaustively acquiring training data per category. This raises the *cross-category problem* between model training and test, i.e. zero-shot samples for unseen categories during training. Such an understanding motivates us to design a cross-modal matching method based on global category-level visual-textual embedding, a common zero-shot learning approach (Fig 2(b)). AAIPR [37] also uses the global embedding idea but totally ignores the zero-shot learning challenge in model design.

As a type of solution for attribute query person search, existing ZSL models are however suboptimal. *First*, unlike the conventional ZSL settings that classify a test image into a small number of categories, we match a text attribute description against massive person images and much more categories. This represents a *larger scale more challenging zero-shot search* problem. Existing state-of-the-art ZSL methods are based on global category-level visual-textual embedding but scale poorly to large tests [35]. A plausible reason is due to insufficient *local attribute-level* discrimination for more fine-grained matching. *Second*, surveillance images in person search present significantly more noise and ambiguity, presenting a more difficult task. *Third*, lacking semantically meaningful person category names prevents exploiting inter-class relationships.

In this study, we formulate a novel *Attribute-Image Hierarchical Matching* (AIHM) method (Fig 2(c)). It performs attribute and image matching for person search at multiple hierarchical levels, including both global category-level visual-textual embedding and local attribute-level feature embedding. This method aims to overcome the limitations of conventional ZSL models and existing text-based person search methods, by benefiting from the generalisation scalability of conventional attribute classification methods. Importantly, cross-modal matching can be end-to-end optimised across all different levels jointly.

Our **contributions** are: **(I)** We formulate *for the first time* an extended ZSL approach to solving a text attribute query person search problem. Our model aims to solve the intrinsic challenge of limited training category data in surveillance videos. **(II)** We propose a novel *Attribute-Image Hierarchical Matching* (AIHM) method. AIHM is able to match more reliably sparse attribute descriptions with noisy surveillance person images at global category and local attribute levels *concurrently*. This goes beyond the common ZSL nearest neighbour search. **(III)** We further introduce a quality-aware fusion scheme for resolving any visual ambiguity problem. Extensive experiments show the superiority of AIHM over the state-of-the-art methods for attribute query person search on three benchmarks: Market-1501 [39], DukeMTMC [27, 23], and PA100K [24].

2. Related Work

Person Search. The most common person search approach is based on taking bounding box images as probes (queries), framed as an extension of the person re-identification problem [11, 21, 39, 17, 22, 7]. However, image queries are not always available in practice. Recently, text query person search has gained increasing attention with search queries as natural language descriptions [20, 19, 4, 3] or short text keywords (text attributes) [37, 16, 32]. These models enable person search on images by verbal or written text descriptions. Using natural language sentences for person search is attractive due to its natural human user friendliness. However, this imposes extra challenges in computational modelling because (1) accurate and rich training data is expensive to obtain, and (2) modelling consistently and reliably rich and complex sentence syntax and its interpretation to arbitrary images is non-trivial, with added difficulties from poor-quality surveillance images. In contrast, short text attribute descriptions offer a more cost-effective and compu-

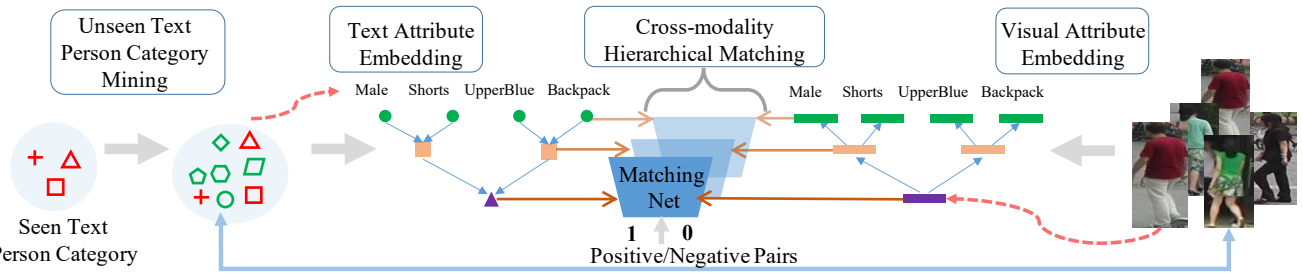

Figure 3: An overview of the proposed *Attribute-Image Hierarchical Matching* (AIHM) model. AIHM is composited of hierarchical visual-textual embedding and cross-modality hierarchical matching. To overcome the one-shot learning challenge in textual embedding, we introduce a simple and effective negative category augmentation strategy in our matching context that allows for enriching the training text data and reducing the model over-fitting risk.

tationally more tractable approach to solving this problem.

Visual Attributes. Computing visual attributes has been extensively used for person search [15, 16, 17, 28, 26, 33, 6]. The idea is to exploit the visual representation of a person by attributes as the mid-level descriptions, which are semantically meaningful and more reliable than low-level pixel feature representations. For example, Peng et al. [26] mine unlabelled latent visual attributes in a limited attribute label space for enriching the appearance representation. Considered as a more domain-invariant or domain adaptive visual feature representation, Wang et al. [33] exploit visual attribute learning for unsupervised identity knowledge transfer across surveillance domains. All these existing methods are focused on visual attribute representations to facilitate image query person search. On the contrary, the focus of this work is on text query person search.

Text Attributes. A few attempts for text attribute query person search have been proposed [32, 16, 37]. In particular, Vaquero et al. [32] and Layne et al. [16] propose the first studies that treat the problem as a multi-label classification learning task. Whilst allowing to flexibly model arbitrary attribute combinations, this strategy has no capacity of modelling the holistic person category information and is therefore suboptimal for processing ambiguous surveillance data. More recently, Yin et al. [37] exploit the idea of cross-modal data alignment. This captures the holistic appearance information of persons, but suffers from a cross-category domain gap problem between the training and test data. In contrast, we uniquely consider the problem from a zero-shot learning perspective and formulate a novel AIHM model. Critically, our model not only addresses the limitation of existing solutions but also combines their modelling merits for enabling extra complementary benefits.

Zero-Shot Learning. Attribute query person search can be understood from zero-shot learning (ZSL) [14, 1, 35, 30, 38], due to the need for generalising to unseen categories in test. But there are several significant differences. First of all, most ZSL methods are designed for image classification other than search/retrieval. The latter is often more challenging due to larger search space. In contrast to conventional ZSL setting, there is no meaningful category names in person search. This disables the exploitation of semantic relationships between seen and unseen categories. Besides, the imagery data of person search often involve more noise and corruption which imposes more difficulty. These factors render the state-of-the-art ZSL methods less effective for person search, as we demonstrate in experiments.

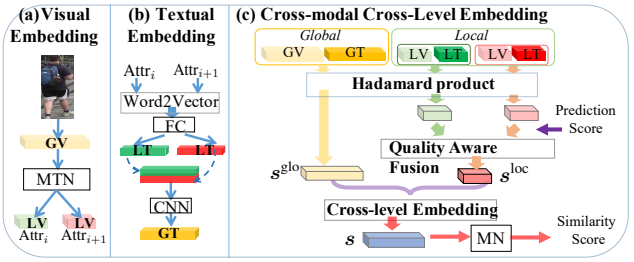

Figure 4: Hierarchical visual-textual embedding and matching. MTN: Multi-Task Network. MN: Matching Net, 3 layer FCs for similarity score prediction.

3. Methodology

To train a textual attribute query person search model, we need to label a set of N image-attribute training pairs $\mathcal{D} = \{\boldsymbol{I}_i, \boldsymbol{a}_i\}_{i=1}^{N}$ describing N_{id} different person descriptions. A multi-label attribute text description of a person image, we call an ***attribute vector*** $\boldsymbol{a}_i$, defines the value of each attribute label with respect to the corresponding person appearance. Persons sharing the same attribute vector description specifying a type of people are considered to belong to a ***person category***. There are a total of N_{att} different binary-class or multi-class attribute labels. We model this problem by zero-shot learning (ZSL) considering that test person categories may be unseen to model training.

3.1. Approach Overview

A schematic overview of the proposed AIHM model is illustrated in Fig 3. The objective of AIHM is to learn a similarity matching model between text attributes $\boldsymbol{a}$ and person

images $\boldsymbol{I}$ in a hierarchical visual-textual embedding space. Instead of nearest neighbour search as most ZSL methods adopt, we aim to learn a similarity matching model: $\hat{y} = f_{\boldsymbol{\theta}}(\boldsymbol{a}, \boldsymbol{I}) \in [0, 1]$, with $\boldsymbol{\theta}$ the model parameters. If a specific text-image pair is a true match, the model should ideally output 1; Otherwise 0. For model training, we adopt the mean square error loss function [30]:

$$\mathcal{L}_{\text{mse}} = \frac{1}{N_{\text{batch}}} \sum_{i=1}^{N_{\text{batch}}} (y_i - \hat{y}_i)^2 \tag{1}$$

where y_i and $\hat{y}_i$ denote the ground-truth and predicted similarity of the i-th training pair, respectively. The mini-batch size is specified by N_{batch}. To enable such matching, we need to form a hierarchical visual-textual embedding (Sec 3.2 & Sec 3.3) and cross-modality fusion (Sec 3.4) as the matching input (Eq (7)). For presentation brevity, in the following we assume a two-level hierarchy: a global category level, and a local per-attribute level. It is straightforward to extend to more hierarchical levels without changing the model designs as described below.

3.2. Hierarchical Visual Embedding

For hierarchical visual embedding of a person image, we employ a multi-task joint learning strategy [5]. An overview of hierarchical visual embedding is given in Fig 4(a). Specifically, we build local attribute-specific embedding ($\boldsymbol{x}_i^{\text{loc}}$, $i \in \{1, \cdots, N_{\text{att}}\}$) based on the global counterpart ($\boldsymbol{x}^{\text{glo}}$) in a ResNet-50 architecture [12]. For each attribute label, we use a separate lightweight branch with two fully connected (FC) layers. The design is suitable since only a small number of ($\sim$10) attributes exist in typical person search scenarios. In cases of many attribute labels, we can assign each branch with a group of attributes for limiting the branch number as well as the overall model complexity (see Table 7 for evaluation).

For discriminative learning of *local attribute-level* visual embedding, we utilise the softmax Cross Entropy (CE) loss. We treat each individual attribute label as a separate classification task ($\mathcal{L}_{\text{cls}}$). Formally, they are formulated as:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N_{\text{batch}}} \sum_{i=1}^{N_{\text{batch}}} \sum_{j=1}^{N_{\text{attr}}} \log(p_{ij}), \tag{2}$$

where p_{ij} is the probability estimation of the i-th training sample on the j-th ground-truth attribute. By multi-task learning, we can obtain the *global category-level* visual embedding as the shared feature representation of all local embeddings. See *supplemental materials* for the network architecture details.

3.3. Hierarchical Textual Embedding

We also need to learn a hierarchical embedding of text attributes. An overview of hierarchical textual embedding is shown in Fig 4(b). Due to small training attribute label data (only one attribute vector per person category), it is challenging to derive a rich textual embedding. In contrast to ZSL, we have *no* access to meaningful person category names in person search. This prevents us from using a wikipedia pre-trained word2vector model to represent person category for benefiting from auxiliary knowledge [25]. For text attributes (also available in person search), the most common representation in ZSL is multi-label binary vector, which however is less effective and informative (Table 6).

To enable the benefit of rich wikipedia information, we propose to represent the attribute labels by word2vector representations. Specifically, we use the word2vector model to map each attribute name into a semantic (300-D) space[1], then further into the *local textual embedding* space $\boldsymbol{z}^{\text{loc}}$ by one FC layer. We then similarly adopt multi-task learning for embedding each attribute label $\boldsymbol{z}_i^{\text{loc}}$, $i \in \{1, \cdots, N_{\text{att}}\}$. To obtain the *global textual embedding* $\boldsymbol{z}^{\text{glo}}$, a simple approach is average pooling per-attribute embeddings. This is likely suboptimal due to lacking of task-specific supervised learning. To overcome this problem, we learn to combine per-attribute embeddings by a *fusion unit* consisting of two 1×1 conv layers. This allows for both intra-attribute and inter-attribute fusion:

$$\boldsymbol{z}^{\text{glo}} = f(\{\boldsymbol{z}_i^{\text{loc}}\}_{i=1}^{N_{\text{att}}}) = \text{Tanh}\Big(\sum_{i=1}^{N_{\text{att}}} \big(\boldsymbol{w}_2^i \cdot \text{Tanh}(\boldsymbol{w}_1^i \cdot \boldsymbol{z}_i^{\text{loc}})\big)\Big), \tag{3}$$

where $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are learnable parameters and Tanh is the non-linear activation function. We use the CE loss functions (Eq (2)) to supervise the textual embedding. In training, the embedding loss and matching loss are *jointly* optimised end-to-end with identical weight. Note, unlike the visual embedding process, we obtain the global category-level textual embeddings by combining all local attribute-level counterparts, an inverse process. This is due to additionally using auxiliary information (wikipedia).

Negative Category Augmentation. The one-shot per category problem in textual modality raises model training difficulty. To alleviate this problem, we exploit negative category augmentation to AIHM model learning. This is achieved by generating new random attribute vectors. We use these synthesised attribute vectors as negative samples in the matching loss (Eq (1)). This helps alleviate the model over-fitting risk whilst enhancing the sparse training data, particularly for global textual embedding. Interestingly, we are not aware of any existing ZSL and person search methods that leverage this simple strategy. One possible reason is that previous methods mostly do not exploit negative

[1]We transform binary attribute labels to binary flags for guaranteed inclusion. Specifically, we transform a binary label "*" as a form of "Yes"+"*" and "No"+"*" before extracting the word2vector label representation. The unknown attribute is set to the vector $\mathbf{0}$.

cross-modality pairs in objective learning loss function. We will verify the efficacy of this scheme (see Fig 6).

3.4. Cross-Modality Cross-Level Embedding

Given hierarchical visual-textual embedding as derived above, we next combine them across modalities and levels to form the final embedding for attribute-image matching. An illustration of this cross-modality cross-level embedding is shown in Fig 4(c). To this end, a common fusion method is *concatenating* two embedding vectors for each training pair [19, 20, 36]. This however may be suboptimal, due to lacking the feature dimension correspondence across modalities which makes the optimisation ineffective. Instead, we deploy *Hadamard Product* that fuses two input vectors by element-wise multiplication.

(I) Cross-Modality Global-Level Embedding. We form the cross-modality global-level embedding $\boldsymbol{s}^{\text{glo}}$ as:

$$\boldsymbol{s}^{\text{glo}} = \boldsymbol{x}^{\text{glo}} \circ \boldsymbol{z}^{\text{glo}}, \tag{4}$$

where $\circ$ specifies the Hadamard product.

(II) Cross-Modality Local-Level Embedding. Unlike the single global-level embedding, we have multiple local per-attribute embeddings in both modalities. Therefore, we first need to form *per-attribute* cross-modality embedding as:

$$\boldsymbol{s}_i^{\text{loc}} = \boldsymbol{x}_i^{\text{loc}} \circ \boldsymbol{z}_i^{\text{loc}}, \quad i \in \{1, \cdots, N_{\text{att}}\}. \tag{5}$$

We then fuse over attributes. Instead of average pooling, we design a ***quality aware fusion*** algorithm. This is based on two considerations: (1) Both surveillance imagery (poor quality with noisy and corrupted observations) and attribute labelling (annotation errors due to poor imaging condition) are not highly reliable. Trusting all attributes and treating them equally in matching are error prone. (2) The significance for person search may vary across attributes.

Specifically, to estimate the per-attribute quality ρ_i^{loc}, we use the minimal prediction scores on image and text as $\rho_i^{\text{loc}} = \min(p_i^{\text{vis}}, p_i^{\text{tex}}), \quad i \in \{1, \cdots, N_{\text{att}}\}$, where p_i^{vis} and p_i^{tex} denote the ground-truth class posterior probability estimated by the corresponding classifier. This discourages the model fit towards corrupted and noisy observations. Based on this quantity measure, we learn a fusion unit (Eq (3)) for adaptively cross-attribute embedding as:

$$\boldsymbol{s}^{\text{loc}} = f\left(\{\rho_i^{\text{loc}} \cdot \boldsymbol{s}_i^{\text{loc}}\}_{i=1}^{N_{\text{att}}}\right). \tag{6}$$

(III) Cross-Modality Cross-Level Embedding. After concatenating the cross-level embeddings, we use a fusion unit (Eq (3)) to form the final cross-modality embedding as:

$$\boldsymbol{s} = f\left(\{\boldsymbol{s}^{\text{loc}}, \boldsymbol{s}^{\text{glo}}\}\right). \tag{7}$$

The final embedding $\boldsymbol{s}$ is used to estimate the attribute-image matching result $\hat{y}$ (Eq (1)) given an input attribute query and person image.

Table 1: Statistics of person search datasets.

Datasets	Market-1501	DukeMTMC	PA100K
# Attribute category	10	8	15
# Train person category	508	300	2020
# Train image	12,936	16,522	80,000
# Test person category	529	387	849
# Unseen	367	229	168
# Test image	15,913	19,889	10,000

4. Experiments

Datasets. In evaluations, we used two publicly available person search (Market-1501 [39], DukeMTMC [27, 23]) and one large pedestrian analysis (PA100K [24]) benchmarks. These datasets present good challenges for person search with varying camera viewing conditions. We followed the standard evaluation setting. The dataset statistics are summarised in Table 1.

Performance Metrics. We used the CMC and mAP as evaluation metrics. As [37], we treated the gallery images respecting a given attribute vector query as true matches.

Implementation Details. For fair comparison to [37], we used ResNet-50 [12] as the backbone net for learning visual embedding. We employed Adam as the optimiser. We set the batch size to 16 (attribute-image pairs), the learning rate to 1e-5, and the epoch number to 150. In each mini-batch, we formed on-the-fly 16/255(16*16-1) positive/negative text-image training pairs. We used 50 training person categories for parameter cross-validation. We used a two-layer hierarchy in AIHM for the main experiments, with different hierarchy structures evaluated independently.

4.1. Comparisons to the State-of-The-Art Methods

Competitors. We compared our AIHM with a wide range of plausible solutions to text attribute person search methods in two paradigms: **(1)** *Global category-level visual-textual embedding* methods: Learning to align the distributions of text attributes and images in a common space, including CCA [2, 34, 8, 29] or MMD [31] based cross-modal matching models, ZSL methods (DEM [38], RN[30], GAZSL [41]), visual semantics embedding (VSE++ [9]), and GAN based cross-modality alignment (AAIPR [37]). **(2)** *Local attribute-level visual-textual embedding* methods: Learning attribute-image region correspondence, including region proposal based dense text-image cross-modal matching (SCAN [18]), natural language query based person search (GAN-RNN [20] and CMCE [19]). We used the officially released codes with careful parameter tuning if needed, e.g. those originally applied to different applications. In testing language models [9, 18, 20, 19], we used random attribute sentences due to no ordering and reported the average results of 10 trials. For all methods, we used ResNet-50 for visual embedding.

Results. The person search performance comparisons on

Table 2: Comparisons to the state-of-the-art methods. **Red**/**Blue**: Best/second best results.

Method	Market-1501				DukeMTMC				PA100K			
	Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10	mAP
DEM[38]	34.0	48.1	57.5	17.0	22.7	43.9	54.5	12.9	20.8	38.7	44.2	14.8
RN[30]	17.2	38.7	47.3	15.5	25.1	42.0	51.5	13.0	**27.5**	38.8	46.6	13.6
GAZSL [41]	23.3	36.9	45.9	14.1	18.2	30.0	37.8	11.9	2.2	3.8	5.3	0.9
DeepCCAE[34]	8.1	23.9	34.5	9.7	33.2	59.3	67.6	14.9	21.2	39.7	48.0	15.6
DeepCCA[2]	29.9	50.7	58.1	17.5	36.7	58.8	65.1	13.5	19.5	40.3	49.0	15.4
2WayNet[8]	11.2	24.3	31.4	7.7	25.2	39.8	45.9	10.1	19.5	26.6	34.5	10.6
MMD[31]	34.1	47.9	57.2	18.9	41.7	**62.3**	68.6	14.2	25.8	38.9	46.2	14.4
DeepCoral[29]	36.5	47.6	55.9	20.0	46.1	61.0	68.1	**17.1**	22.0	39.7	48.1	14.1
VSE++[9]	27.0	49.1	58.2	17.2	33.6	54.7	62.8	15.5	22.7	39.8	48.1	**15.7**
AAIPR[37]	**40.2**	49.2	**58.6**	20.6	**46.6**	59.6	**69.0**	15.6	27.3	**40.5**	**49.8**	15.2
SCAN[18]	4.0	10.1	15.3	2.1	3.5	9.3	14.3	1.6	2.9	8.2	12.5	1.9
GNA-RNN[20]	30.4	38.7	44.4	15.4	34.6	52.7	65.8	14.2	20.3	30.8	38.2	9.3
CMCE[19]	35.0	**50.9**	56.4	**22.8**	39.7	56.3	62.7	15.4	25.8	34.9	45.4	13.1
AIHM	**43.3**	**56.7**	**64.5**	**24.3**	**50.5**	**65.2**	**75.3**	**17.4**	**31.3**	**45.1**	**51.0**	**17.0**

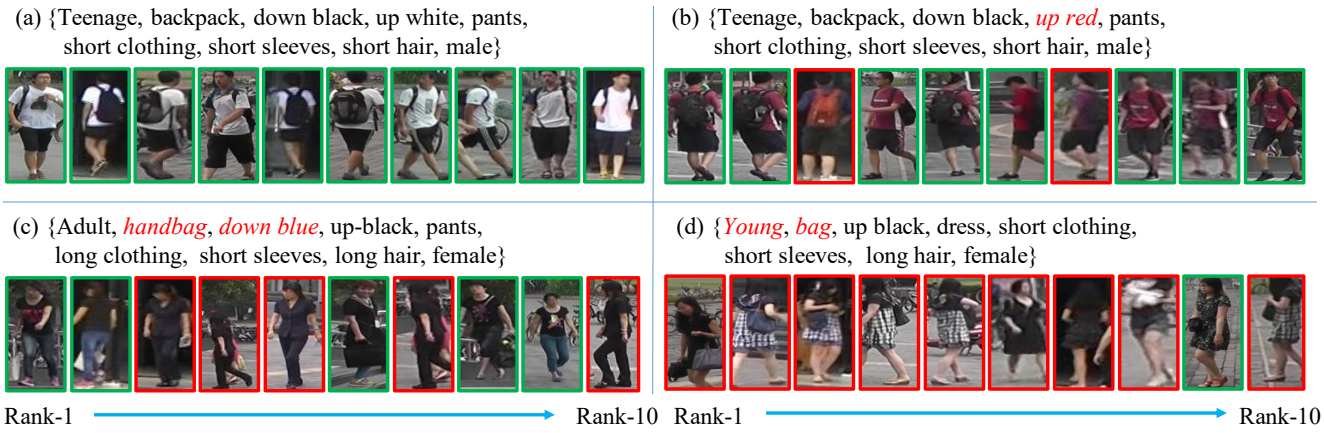

Figure 5: Examples of person search by attribute query on Market-1501. Attribute query is on the top in each case. True/false image matches are indicated by green/red boxes. We highlight the attributes in red corresponding to the false matches.

three benchmarks are shown in Table 2. It is evident that our AIHM model outperforms all the existing methods, e.g. surpassing the second best and state-of-the-art person search model AAIPR [37] by a margin of 3.1%/3.7% in Rank-1/mAP on Market-1501. The performance margins over other global visual-textual embedding methods and local region correspondence learning model are even more significant. In particular, state-of-the-art ZSL models also fail to excel due to the larger scale search, more ambiguous visual observation, and meaningless category names. Overall, these results show that despite their respective modelling strength either global and local embedding *alone* is suboptimal for the more challenging person search problem. It is clearly beneficial to the overall model performance if their complementary advantages are utilised as formulated in the AIHM model.

4.2. Qualitative Analysis and Visual Examination

To provide more in-depth and visual examination on the performance of AIHM, we conducted a qualitative analysis, as shown in Fig 5. It is clear that the majority of the search results in top-10 by AIHM match the attribute query precisely, with a few false matches due to the very similar visual appearance of different person categories. For example, AIHM succeeds in detecting the tiny "handbag" in the Rank1 image (c) and the "backpack" with the very limited visible part in the Rank1 image (a), thanks to the capability of local correspondence matching across modalities.

We found that false retrieval images are often due to ambiguous visual appearances and/or text descriptions. For example, the Rank7 image (b) is with "up-purple" whilst the Rank9 with "up-red". Such a colour difference is visually very subtle even for humans. Another example with visual

ambiguity is "blue" vs "black" (c). In terms of ambiguous text attribute descriptions "Teenage" and "Young" are semantically very close. This causes the failure search results (d), where "Teenage" person images in top-7 are instead retrieved against the query attribute "Young".

4.3. Further Analysis and Discussion

Hierarchical embedding and matching. We examined the effect and complementary of joint local attribute-level and global category-level visual-textual embedding in AIHM. This is conducted by comparing individual performances with their combinations. Table 3 shows that: (1) Either embedding *alone* is already considerably strong and discriminative for person search. Local AIHM embedding alone is competitive to the state-of-the-art AAIPR [37]. (2) A clear performance gain is obtained by combining both global and local embedding as a whole in person search. This validates the complementary benefits and performance advantages of jointly learning local and global visual-textual embedding interactively in AIHM.

Table 3: Hierarchical embedding and matching analysis.

Method	Market-1501		DukeMTMC		PA100K	
	Rank1	mAP	Rank1	mAP	Rank1	mAP
Global Only	30.6	20.5	40.7	13.7	26.1	14.3
Local Only	39.5	21.9	46.9	15.3	29.4	15.6
Hierarchy	**43.3**	**24.3**	**50.5**	**17.4**	**31.3**	**17.0**

Quality-aware fusion. Recall that we included quality-aware fusion (Eq (6)) in AIHM for alleviating the negative effect of noisy and ambiguous observation in local visual-textual embedding. We tested the efficacy of this component in comparison to the common average pooling strategy. Table 4 shows that our quality-aware fusion is more effective in suppressing noisy information, e.g. improving over the average pooling in Rank1/mAP rates by 4.3%/0.5% on Market-1501, 5.6%/1.3% on DukeMTMC, and 5.2%/1.9% on PA100K, respectively. This shows the benefit of taking into account the input data quality in person search.

Table 4: Quality-aware fusion vs. Average Pooling.

Method	Market-1501		DukeMTMC		PA100K	
	Rank1	mAP	Rank1	mAP	Rank1	mAP
Avg Pool	39.0	23.8	44.9	16.1	26.1	15.1
AIHM	**43.3**	**24.3**	**50.5**	**17.4**	**31.3**	**17.0**

Negative category augmentation. To combat the one-shot learning challenge in global textual embedding, we exploited negative category augmentation in AIHM model learning, so to enrich training text data for reducing overfitting risk. We tested three different augmentation sizes: 5k, 10k, and 20k. It is shown in Fig 6 that this text augmentation is clearly beneficial to AIHM. For example, with 10k negative categories, we obtained 4.4%, 5.5% and 3.8% gain at Rank-1 on Market-1501, DukeMTMC, and PA100K, respectively. The optimal augmentation size is around 10k. Its benefit can be understood from a negative hard mining viewpoint, which improves model discriminative learning given limited training category data. However, too many (e.g. 20k) negative pairs seem to have negatively overwhelmed model learning due to limited positive pairs.

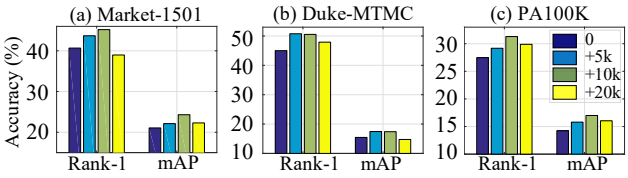

Figure 6: Text negative category augmentation.

Table 5: Model design strategy examination: Attribute Recognition (AR) vs Learning to Compare (as AIHM).

Dataset	Methods	Rank1	Rank5	Rank10	mAP
Market-1501	AR	35.7	47.8	57.8	19.8
	AIHM	**43.3**	**56.7**	**64.5**	**24.3**
DukeMTMC	AR	42.0	52.9	63.2	15.8
	AIHM	**50.5**	**65.2**	**75.3**	**17.4**
PA100K	AR	30.3	42.8	47.8	13.8
	AIHM	**31.3**	**45.1**	**51.0**	**17.0**

Person search by individual attribute recognition. We examined two high-level model design strategies for person search: **(1)** Attribute Recognition (AR): Using the attribute prediction scores by the AIHM's visual component, and the L_2 distance metric in the attribute vector space for cross-modal matching and ranking. **(2)** Learning to match strategy, i.e. the AIHM, which considers both global category-level and local attribute-level textual-visual embedding. It is interesting to find from Table 5 that the AR baseline performs reasonably well when compared to the competitors in Table 2. For example, AR even approaches the performance of the state-of-the-art person search model AAIPR [37]. Note that, this strong AR is likely to benefit from our hierarchical embedding learning design. Besides, the big performance margins of our model over AR suggest that the learning to match strategy in joint optimisation is superior.

Global textual embedding. We examined three design considerations for learning the global textual embedding: **(1)** Individual attribute representation: One-Hot (OH) vs Word2Vec (WV), **(2)** Aggregation of multiple attribute embedding: RNN (LSTM) vs CNN. **(3)** Binary-class label representation: Zero vs Transformed Input. Table 6 shows that:

(1) OH+CNN outperforms OH+RNN, suggesting that artificially introducing the modelling of temporal structure information on *orderless* person attributes is not only unnecessary but also brings adverse effect to model performance. (2) WV+CNN outperforms OH+CNN, indicating that WV is a more informative attribute representation particularly in case of sparse training attribute data. Our textual embedding design via CNN is superior to directly using WV, suggesting the necessity of feature transformation because the generic WV is not optimised particularly for person image analysis.

Table 6: Global textual embedding analysis. OH: One-Hot; WV: Word2Vec.

Method	Market-1501		DukeMTMC		PA100K	
	Rank1	mAP	Rank1	mAP	Rank1	mAP
OH+RNN	35.7	17.8	46.6	16.8	21.4	12.3
OH+CNN	**37.1**	**21.0**	**49.8**	**18.1**	**25.3**	**13.7**
WV	41.8	22.9	48.7	16.2	29.1	14.2
OH+CNN	39.1	22.0	46.5	16.1	25.3	13.7
WV+CNN	**43.3**	**24.3**	**50.5**	**17.4**	**31.3**	**17.0**

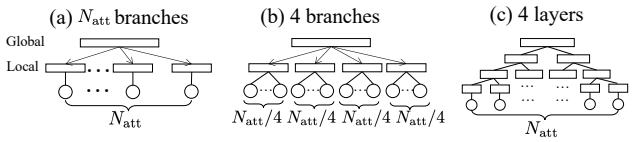

Figure 7: Hierarchy variants. (a) Two levels, one branch one attribute, N_{att} branches totally; (b) Two levels, one branch $N_{att}/4$ attributes, 4 branches totally; (c) Four levels, 2 branches at layer-2, end by N_{att} branches.

Multi-task learning scalability. We use multi-task learning for local visual-textual embedding, so the branch number is decided by the attribute set size N_{att} (Fig. 7 (a)). For scaling to cases of many attributes, we can use a branch for a group of attributes. We conducted a controlled evaluation with two hierarchical layers. Given N_{att} attributes, we randomly grouped them into 4 size-balanced groups before applying AIHM (Fig. (b)). We repeated 5 trials of different grouping and reported the average results. Table 7 shows that attribute grouping reduces model performance due to less fine-grained local embedding, as expected. Importantly, the performance drop is not significant. This also verifies our AIHM design motivation of incorporating local and global embedding jointly, in contrast to state-of-the-art ZSL methods that consider global embedding alone.

Hierarchy depth. We evaluated the effect of AIHM's hierarchy depth on model performance. We used random grouping to form size-balanced intermediate layers for l-layers ($l = 2/4$) hierarchies (see Fig. 7(c)). The results were averaged over 5 trials. Table 8 shows that a hierarchy with more layers leads to better model performance but come with higher computational costs (one feature vector per hierarchy node per modality, fusion over all layers).

Table 7: Scalability of multi-task learning local embedding.

#Branch	Market-1501		DukeMTMC		PA100K	
	Rank1	mAP	Rank1	mAP	Rank1	mAP
$N_{att}/4$	41.6	23.9	47.9	15.6	30.3	16.3
N_{att}	**43.3**	**24.3**	**50.5**	**17.4**	**31.3**	**17.0**

Table 8: Effect of hierarchy depth.

#Depth	Market-1501		DukeMTMC		PA100K	
	Rank1	mAP	Rank1	mAP	Rank1	mAP
2	43.3	24.3	50.5	17.4	31.3	17.0
4	**45.2**	**25.2**	**53.6**	**18.5**	**33.4**	**17.8**

5. Conclusion

In this work, we presented a novel *Attribute-Image Hierarchical Matching* (AIHM) model for text attribute query person search. Unlike most existing methods, which assume image based queries that are not always available in practice, AIHM enables person search with only short text attribute descriptions. In contrast to few existing methods for attribute query person search, we formulate this problem as an extended zero-shot learning problem with a more principled approach to its solution. Algorithmically, the proposed AIHM model solves the fundamental limitations of existing ZSL learning methods by joint global category-level and local attribute-level visual-textual embedding and matching. This aims to eliminate their respective modelling weaknesses whilst optimising their mutual complementary advantages. Extensive comparative evaluations demonstrated the performance superiority of the proposed AIHM model over a wide range of existing alternative methods on three attribute person search benchmarks. We provided detailed component analysis for giving insights on model design and its performance advantages.

Acknowledgements

This work is supported by Vision Semantics Limited, the China Scholarship Council, the Alan Turing Institute, and Innovate UK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety (98111-571149).

References

[1] Ziad Al-Halah and Rainer Stiefelhagen. How to transfer? zero-shot object recognition via hierarchical transfer of semantic attributes. In *WACV*, 2015. 3

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML*, 2013. 5, 6

[3] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *ECCV*, 2018. 2

[4] Tianlang Chen, Chenliang Xu, and Jiebo Luo. Improving text-based person search by spatial matching and adaptive threshold. In *WACV*. 2

[5] Qi Dong, Shaogang Gong, and Xiatian Zhu. Multi-task curriculum transfer deep learning of clothing attributes. In *WACV*, 2017. 4

[6] Qi Dong, Shaogang Gong, and Xiatian Zhu. Imbalanced deep learning by minority class incremental rectification. *TPAMI*, 2018. 3

[7] Qi Dong, Xiatian Zhu, and Shaogang Gong. Single-label multi-class image classification by deep logistic regression. In *AAAI*, 2019. 2

[8] Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *CVPR*, 2017. 5, 6

[9] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. 2018. 5, 6

[10] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-view zero-shot learning. *TPAMI*, 2015. 2

[11] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person re-identification*. Springer, 2014. 1, 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 5

[13] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 1

[14] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *TPAMI*, 2014. 3

[15] Ryan Layne, Timothy M Hospedales, and Shaogang Gong. Person re-identification by attributes. In *BMVC*, 2012. 3

[16] Ryan Layne, Timothy M Hospedales, and Shaogang Gong. Attributes-based re-identification. In *Person Re-Identification*. Springer, 2014. 1, 2, 3

[17] Ryan Layne, Timothy M Hospedales, and Shaogang Gong. Re-id: Hunting attributes in the wild. In *BMVC*, 2014. 2, 3

[18] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, 2018. 5, 6

[19] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *ICCV*, 2017. 1, 2, 5, 6

[20] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. *CVPR*, 2017. 1, 2, 5, 6

[21] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 1, 2

[22] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 2

[23] Yutian lin, Liang Zheng, and Wu Yu and Yang Yi Zheng, Zhedong and. Improving person re-identification by attribute and identity learning. *arXiv*, 2017. 2, 5

[24] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, 2017. 2, 5

[25] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 4

[26] Peixi Peng, Yonghong Tian, Tao Xiang, Yaowei Wang, Massimiliano Pontil, and Tiejun Huang. Joint semantic and latent attribute modelling for cross-class transfer learning. *TPAMI*, 2018. 3

[27] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV workshop on Benchmarking Multi-Target Tracking*, 2016. 2, 5

[28] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *ECCV*, 2016. 3

[29] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, 2016. 5, 6

[30] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018. 3, 4, 5, 6

[31] Ilya O Tolstikhin, Bharath K Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximum mean discrepancy with radial kernels. In *NIPS*. 5, 6

[32] Daniel A Vaquero, Rogerio S Feris, Duan Tran, Lisa Brown, Arun Hampapur, and Matthew Turk. Attribute-based people search in surveillance environments. In *Workshop of WACV*, 2009. 1, 2, 3

[33] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR*, 2018. 3

[34] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *ICML*, 2015. 5, 6

[35] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 2018. 2, 3

[36] Weidi Xie, Li Shen, and Andrew Zisserman. Comparator networks. *ECCV*, 2018. 5

[37] Zhou Yin, Wei-Shi Zheng, Ancong Wu, Hong-Xing Yu, Hai Wan, Xiaowei Guo, Feiyue Huang, and Jianhuang Lai. Adversarial attribute-image person re-identification. In *IJCAI*, 2018. 1, 2, 3, 5, 6, 7

[38] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2017. 3, 5, 6

[39] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 1, 2, 5

[40] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *TPAMI*, 2013. 1

[41] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*, 2018. 5, 6